INFORMATION AND CONTROL 39, 227–242 (1978)

f-Entropies, Probability of Error, and Feature Selection*

MOSHE BEN-BASSAT

Center for the Critically Ill, School of Medicine, University of Southern California, 1300 North Vermont Avenue, Los Angeles, California

The f-entropy family of information measures: $u(p_1,\ldots,p_m) = \sum f(p_k)$, f concave (e.g., Shannon (1948) *Bell Syst. Tech. J.* **27**, 379–423, 623–656; Suadratic; Daroczy (1970) *Inform. Contr.* **16**, 36–51; etc.), is considered. Characterization of the tightest upper and lower bounds on f-entropies by means of the probability of error, is presented. These bounds are used to derive the dual bounds, i.e., the tightest lower and upper bounds on the probability of error by means of f-entropies. Concerning the use of f-entropies as a tool for feature selection, it is proved that none of the members of this family induce over an arbitrary set of features the same preference order as does the probability of error rule.

1. INTRODUCTION

Consider the classical pattern recognition problem which is concerned with the assignment of a given object to one of m known classes. The uncertainty about the true class is expressed by the prior probability vector $\Pi = (\pi_1 \cdots \pi_m)$, where π_i represents the prior probability of class i, $\pi_i \geqslant 0$ and $\sum_{i=1}^{m} \pi_i = 1$. This uncertainty can be modified by observing features of the object to be classified. Let F denote the set of all observable features, each of which is represented by a real random variable, which in some cases may be multidimensional. In what follows, X, Y, Z denote random variables in F. Adopting the Bayesian approach, the true class is considered as a real random variable C taking values in the finite set $\{1, 2,\ldots, m\}$.

Once a value x is observed for $X \in F$, the posterior probability of class i, $P(C = i)$, is given by Bayes' theorem:

$$\pi_i(x) = \frac{\pi_i f_{iX}(x)}{\sum_{k=1}^{m} \pi_k f_{kX}(x)}, \qquad i = 1, 2,\ldots, m, \tag{1}$$

where f_{iX} is the conditional generalized density function of X given class i, i.e., f_{iX} is a density function for continuous X and a probability function for discrete

* This study was supported by United States Public Health Service Research Grant RO1HS01474 from the Health Resources Administration.

0019-9958/78/0393-0227$02.00/0

X. Since the denominator vanishes with probability zero, the definition of $\pi_i(x)$ on this null set is irrelevant.

It is well known (Ferguson, 1967) that the decision rule which minimizes the probability of error is the Bayes' decision rule which assigns x to the class with the highest a posteriori probability. Using this rule, the probability of error for a given x is expressed by

$$P(e \mid x) = 1 - \max\{\pi_1(x),..., \pi_m(x)\}. \tag{2}$$

Prior to observing X, the probability of error, $P_X(e)$, associated with X is defined as the expected probability of error after observing it:

$$P_X(e) = E_X[1 - \max\{\pi_1(x) \cdots \pi_m(x)\}]. \tag{3}$$

The approach to feature selection which is adopted in this paper is that of choosing $X \in F$ for which $P_X(e)$ is minimized.

Unfortunately, computing $P_X(e)$ is often impractical; particularly in the multidimensional case and thus, a substitute rule for feature selection is of great importance. An ideal rule is one which for very prior probability vector induces over F the same selection preference as does the probability of error rule. For $m = 2$ and Gaussian features, a partial solution is provided by using the Kullback divergence measure (Marill and Green, 1963; Fu, *et al.*, 1970). However, for an arbitrary set of features and $m \geqslant 2$, no rule is known with the above property.

Since "ideal rules" cannot be obtained, the assessment of a feature selection criterion can be made by considering the tightness and the rate of change of lower and upper bounds on the probability of error derived from this criterion. This approach was adopted by most of the previous investigators (see, for instance, Kailath, 1967, Chen, 1971; Lissack and Fu, 1976).

Another useful tool is provided by examining the dual bounds, i.e., the lower and upper bounds on the criterion function by means of the probability of error. Such bounds were developed for Shannon's entropy by Kovalevski (1968) and their use for feature selection is discussed by Chen (1971).

The purpose of this paper is to extend the results obtained by Kovalevski to a general family of information measures and to use these results for proving that none of the members in this family can serve as an ideal rule in the above sense.

2. Mathematical Framework for Feature Selection

Definition 1. Let X and Y be features in F. X is said to be not preferred to Y, $Y \gtrsim X$, if $P_X(e) \geqslant P_Y(e)$.

Obviously, $\gtrsim$ is a complete order relation (transitive, reflexive, and defined

for every pair of features in F) which induces over F the following preference ($\succ$) and indifference ($\sim$) relations:

$$Y \succ X \quad \text{if} \quad Y \gtrsim X \quad \text{but not} \quad X \gtrsim Y, \tag{4}$$

$$Y \sim X \quad \text{if} \quad Y \gtrsim X \quad \text{and} \quad X \gtrsim Y. \tag{5}$$

The relation $\sim$ is easily seen to be reflexive, symmetric, and transitive and thus, it is an equivalence relation which divides F into equivalent classes defined by:

$$F(p_e) = \{X \mid X \in F, P_X(e) = p_e\}. \tag{6}$$

Denote by R the set of all real numbers and denote

$$A^m = \left\{\Pi \mid \Pi = (\pi_1, \pi_2 \cdots \pi_m), \pi_i \geqslant 0, \sum_{i=1}^{m} \pi_i = 1\right\}, \qquad m \geqslant 2. \tag{7}$$

For a real-valued function u on A^m, a given prior probability vector and a feature X, $X \in F$, denote,

$$U(X) = E[u(\Pi(X) \mid \Pi], \tag{8}$$

where the expectation is taken with respect to the mixed distribution of X. For simplicity of the notation, the dependence on Π is not expressed in $U(X)$.

Comment. In this paper the ordering of features is considered with respect to a fixed set of prior probabilities. Blackwell (1951) proposed to consider X as preferred to Y if $P_X(e) < P_Y(e)$ for all prior distributions on the underlying classes. For the two class case, Blackwell also presented a necessary and sufficient condition for this preference relation by means of the likelihood ratios under the two features.

The notion of "sufficient experiments", as proposed by Blackwell (1951, 1953), is a more general tool for identifying features which are uniformly dominated by other features. The practical meaning of sufficiency is as follows: Y is sufficient for X if, regardless of the true value of C, performing X is equivalent to performing Y and then subjecting the outcome y to a random transformation dominated by a known density function. DeGroot (1970) proved that if Y is sufficient for X then Y is at least as informative as X for all possible a priori probabilities and whatever the true class is. An information function for this purpose is a real valued nonnegative concave function on A^m.

DEFINITION 2. A real-valued function u on A^m, $m \geqslant 2$, is said to represent the relation $\gtrsim$ if:

$$Y \gtrsim X \text{ implies } U(X) \geqslant U(Y) \tag{9}$$

for every X, $Y \in F$, and for every prior probability vector $\Pi \in A^m$.

Due to the completeness of the relation $\gtrsim$, a function u which represents $\gtrsim$ also satisfies

$$Y \succ X \text{ implies } U(X) > U(Y), \tag{10}$$

$$Y \sim X \text{ implies } U(X) = U(Y). \tag{11}$$

By definition of the relation $\gtrsim$, the function u_e defined by

$$u_e(\Pi) = 1 - \max\{\pi_1, ..., \pi_m\}, \qquad \Pi \in A^m, \quad m \geqslant 2, \tag{12}$$

represents this relation.

A popular family in which alternative representative functions for $\gtrsim$ were explored is the family of information measures T defined by

$$T = \Big\{ u \mid u\colon A^m \to R,\ m \geqslant 2,\ u = \sum_{i=1}^{m} f(\pi_i),\ f \text{ strictly concave},$$
$$f'' \text{ exists}, f(0) = \lim_{\pi \to 0} f(\pi) = 0 \Big\} \tag{13}$$

Some of the members in this family are:

(a) Shannon's (1949) entropy

$$u(\Pi) = -\sum_{i=1}^{m} \pi_i \log \pi_i \tag{14}$$

for which

$$f(\pi) = -\pi \log \pi \tag{15}$$

(If not otherwise stated, the base 2 logarithm is assumed.)

(b) Quadratic entropy

$$u(\Pi) = \sum \pi_i(1 - \pi_i), \tag{16}$$

for which

$$f(\pi) = \pi(1 - \pi). \tag{17}$$

The function $u(\Pi)$ first appeared in the context of risk evaluation for the nearest neighbor classification rule (Cover and Hart, 1967). The term quadratic entropy was coined by Vajda (1968). Ito (1972) and Devijver (1973) also analyzed this function.

(c) Darocry's (1970) entropy

$$u(\Pi) = \frac{\sum \pi_i^{\alpha} - 1}{2^{1-\alpha} - 1} \qquad \alpha \neq 1 \tag{18}$$

for which

$$f(\pi) = \frac{\pi^\alpha - \pi}{2^{1-\alpha} - 1} \qquad \alpha \neq 1. \tag{19}$$

General properties of divergence measures, derived from this family, were investigated by Csizar (1963). Following his terminology this family will be named f-entropy measures.

An example of an entropy function which does not belong to the f-entropy family is Renyi's, (1960) entropy of order α, which is given by:

$$H\alpha(\Pi) = (1 - \alpha)^{-1} \log \sum \pi_i{}^\alpha, \qquad \alpha > 0, \quad \alpha \neq 1.$$

The relationships between Renyi's entropy and the probability of error are discussed by Ben-Bassat and Raviv (1976, 1978) and by Toussaint (1977).

In what follows, although we assume that all the features in F have known conditional distributions, these conditional distributions are not limited in any manner. For instance, the values for the probability of error which may be attained under these distributions are not limited to a certain subset of the $[0, 1 - 1/m]$ interval.

3. Bounds on f-Entropies by Means of the Probability of Error

In this section, relationships are derived between the probability of error and the information measures included in T. These relationships are summarized in four theorems, which have already been proved by Kovalevsky (1968) for the special case, $u(\pi) = -\pi \log \pi$. Following Kovalevsky's lines, no difficulties arise in extending his results to a general u, $u \in T$, and therefore, the proofs are omitted here.

Denote

$$A^m(\pi_e) = \{\Pi \mid \Pi \in A^m,\ 1 - \max\{\pi_1 \cdots \pi_m\} = \pi_e\}, \tag{20}$$

$$\bar{u}(\pi_e) = \sup_{\Pi}\{u(\Pi) \mid \Pi \in A^m(\pi_e)\}, \tag{21}$$

$$\underline{u}(\pi_e) = \inf_{X}\{u(\Pi) \mid \Pi \in A^m(\pi_e)\}, \tag{22}$$

$$\bar{U}(p_e) = \sup_{X}\{U(X) \mid X \in F,\ P_X(e) = p_e\}, \tag{23}$$

$$\underline{U}(p_e) = \inf_{Y}\{U(X) \mid X \in F,\ P_X(e) = p_e\}. \tag{24}$$

Theorem 1. *For every $u \in T$ and for any given π_e, $0 \leqslant \pi_e \leqslant 1 - 1/m$, $m \geqslant 2$,*

$\bar{u}(\pi_e)$ is attained when all the components of Π, except perhaps one, are equal to $\pi_e/(m-1)$, and then,

$$\bar{u}(\pi_e) = f(1-\pi_e) + (m-1)f\left(\frac{\pi_e}{m-1}\right). \tag{25}$$

THEOREM 2. *For every $u \in T$ and for any given π_e, $0 \leqslant \pi_e \leqslant 1 - 1/m$, $m \geqslant 2$, $\underline{u}(\pi_e)$ is attained when all the components of Π, except perhaps one, are equal either to $1-\pi_e$ or to zero, and then,*

$$\underline{u}(\pi_e) = tf(1-\pi_e) + f[1 - t(1-\pi_e)], \tag{26}$$

where t is an integer determined by the inequalities

$$t \leqslant \frac{1}{1-\pi_e} < t+1 \tag{27}$$

or equivalently

$$\frac{t-1}{t} \leqslant \pi_e < \frac{t}{t+1}. \tag{28}$$

LEMMA 3. *For every $u \in T$, $\bar{u}(\pi_e)$ is a concave function over the interval $[0, 1 - 1/m]$, $m \geqslant 2$.*

Proof. Let π_e, π_e' be points in $[0, 1-1/m]$, and let λ be a real number $0 \leqslant \lambda \leqslant 1$. By Theorem 1 and the strict concavity of f it follows that for a given $u \in T$,

$$\begin{aligned}
&\bar{u}(\lambda\pi_e + (1-\lambda)\pi_e') \\
&\quad = f(1 - \lambda\pi_e - (1-\lambda)\pi_e') + (m-1)f\left(\frac{\lambda\pi_e + (1-\lambda)\pi_e'}{m-1}\right) \\
&\quad = f(\lambda(1-\pi_e) + (1-\lambda)(1-\pi_e')) \\
&\qquad + (m-1)f\left(\lambda\frac{\pi_e}{m-1} + (1-\lambda)\frac{\pi_e'}{m-1}\right) \\
&\quad > \lambda f(1-\pi_e) + (1-\lambda)f(1-\pi_e') + (m-1)\lambda f\left(\frac{\pi_e}{m-1}\right) \\
&\qquad + (m-1)(1-\lambda)f\left(\frac{\pi_e'}{m-1}\right) \\
&\quad = \lambda\bar{u}(\pi_e) + (1-\lambda)\bar{u}(\pi_e').
\end{aligned}$$

Q.E.D.

LEMMA 4. *For every $u \in T$ and for any given integer t, $t \leqslant m-1$, $m \geqslant 2$, $u(\pi_e)$ is strictly concave over the interval $[(t-1)/t, t/(t+1))$.*

Proof. Consider π_e, π'_e both in the interval $[(t-1)/t, t/(t+1))$ for t integer, and let λ be a real number $0 \leqslant \lambda \leqslant 1$. By Theorem 2 and the strict concavity of f it follows that for a given $u \in T$,

$$\begin{aligned}
&\underline{u}(\lambda\pi_e + (1-\lambda)\pi'_e) \\
&\quad = tf(1 - \lambda\pi_e - (1-\lambda)\pi'_e + f[1 - t(1 - \lambda\pi_e - (1-\lambda)\pi'_e)] \\
&\quad = tf(\lambda(1-\pi_e) + (1-\lambda)(1-\pi'_e) \\
&\qquad + f[\lambda(1 - t(1-\pi_e)) + (1-\lambda)(1 - t(1-\pi'_e))] \\
&\quad > \lambda tf(1-\pi_e) + (1-\lambda)\, tf(1-\pi'_e) + \lambda f[1 - t(1-\pi_e)] \\
&\qquad + (1-\lambda) f[1 - t(1-\pi'_e)] \\
&\quad = \lambda\bar{u}(\pi_e) + (1-\lambda)\,\underline{u}(\pi'_e). \qquad \text{Q.E.D.}
\end{aligned}$$

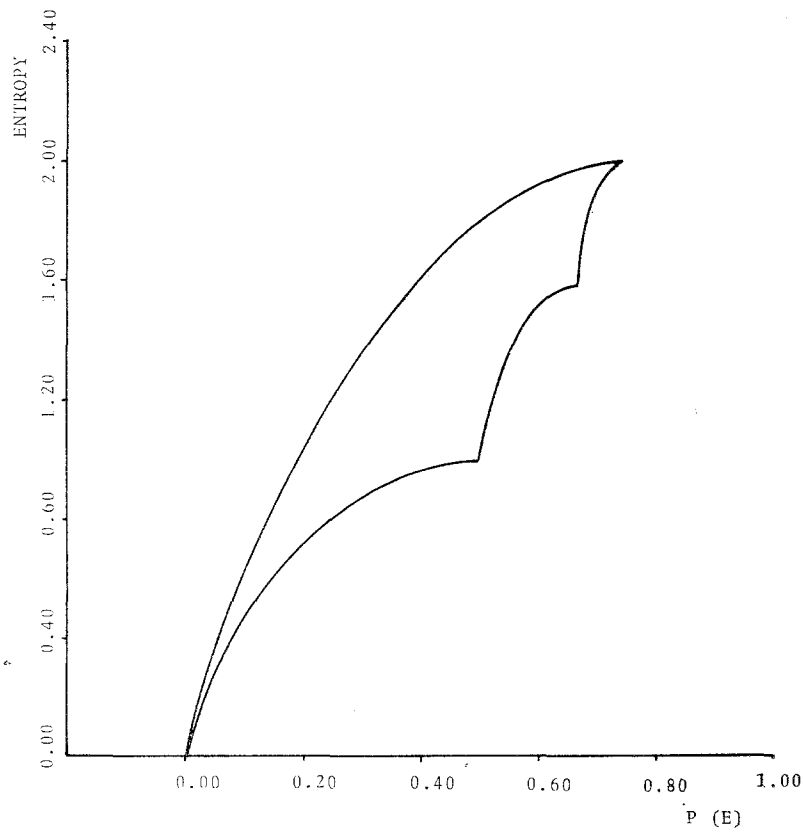

FIG. 1. Graphs of $\bar{u}(p_e)$, $\underline{u}(p_e)$ for Shannon's entropy ($m = 4$).

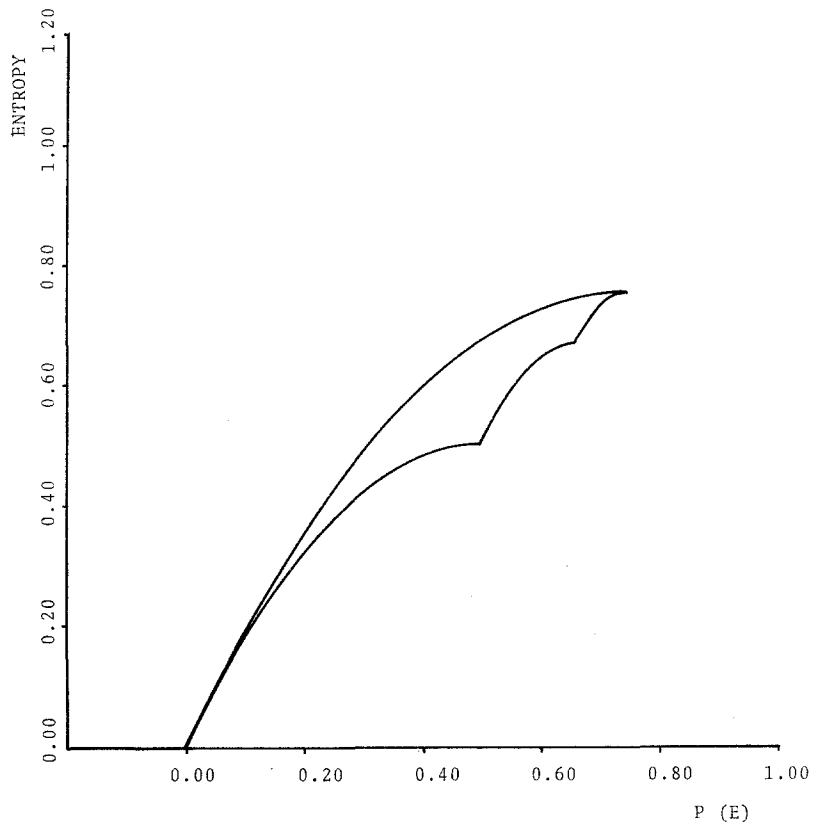

FIG. 2. Graphs of $\bar{u}(p_e)$, $\underline{u}(p_e)$ for the quadratic entropy ($m = 4$).

The significance of the last lemma is that $\underline{u}(\pi_e)$ is piecewise concave over partial intervals of $[0, 1 - 1/m]$. However, on the whole range $[0, 1 - 1/m]$, $\underline{u}(\pi_e)$ is not necessarily concave as shown in Figs. 1 and 2 which portrays $\underline{u}(\pi_e)$ and $\bar{u}(\pi_e)$ for two members of T.

For future use let us construct and discuss the piecewise linear function $L(\pi_e)$ which connects the points $\underline{u}((t-1)/t)$ and $\underline{u}(t/(t+1))$ for $t = 1, 2, \ldots, m-1$ (see Fig. 3)

$$\begin{aligned} L(\pi_e) &= \underline{u}\left(\frac{t-1}{t}\right) + \frac{\underline{u}(t/(t+1)) - \underline{u}((t-1)/t)}{t/(t+1) - (t-1)/t}\left(\pi_e - \frac{t-1}{t}\right) \\ &= tf\left(\frac{1}{t}\right) + t(t+1)\left[(t+1)f\left(\frac{1}{t+1}\right) - tf\left(\frac{1}{t}\right)\right]\left(\pi_e - \frac{t-1}{t}\right). \end{aligned} \tag{29}$$

LEMMA 5. *For every $u \in T$, $L(\pi_e)$ is a convex function.*

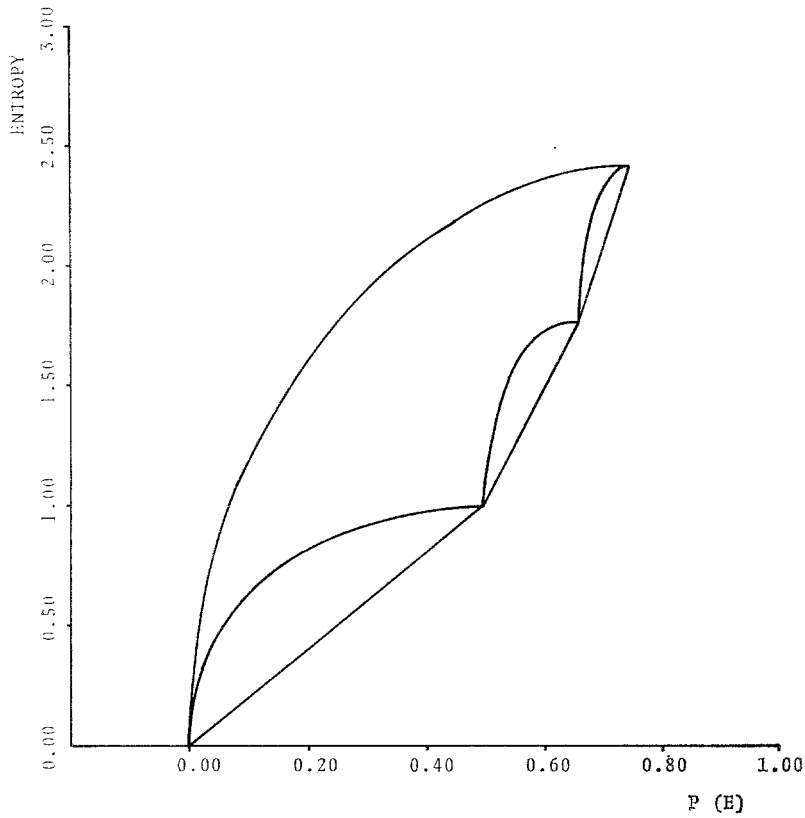

FIG. 3. Graphs of $\bar{u}(p_e)$, $\underline{u}(p_e)$, and $L(p_e)$ for Daroczy's entropy, $\alpha = 0.5$ ($m = 4$).

Proof. A piecewise linear function is convex if its slope never decreases. Consider the slope of $L(\pi_e)$ without the positive factor $t(t+1)$,

$$a(t) = (t+1)f\left(\frac{1}{t+1}\right) - tf\left(\frac{1}{t}\right), \tag{30}$$

$$a'(t) = \left[f\left(\frac{1}{t+1}\right) - \frac{1}{t+1}f'\left(\frac{1}{t+1}\right)\right] - \left[f\left(\frac{1}{t}\right) - \frac{1}{t}f'\left(\frac{1}{t}\right)\right]. \tag{31}$$

Taking the derivative of $b(t) = f(1/t) - (1/t)f'(1/t)$ we obtain $b'(t) = (1/t^3)f''(1/t)$ which is nonpositive for positive t and concave f. Hence, $b(t)$ is a decreasing function of t and therefore $a'(t)$ is nonnegative for positive t, which implies that the slope $a(t)$ never decreases. Q.E.D.

Theorems 1 and 2, respectively, provide upper and lower bounds on the partial information for a given partial probability of error. The next two theorems

provide the respective bounds for the expected information and the expected probability of error.

THEOREM 6. *For every $u \in T$ and for any given value p_e, $0 \leqslant p_e \leqslant 1 - 1/m$, $m \geqslant 2$,*

$$\overline{U}(p_e) = \bar{u}(p_e). \tag{32}$$

$\overline{U}(p_e)$ is attained when all the partial error probabilities are equal to p_e.

THEOREM 7. *For every $u \in T$ and for any given values p_e, $0 \leqslant p_e \leqslant 1 - 1/m$, $m \geqslant 2$,*

$$\underline{U}(p_e) \leqslant L(p_e). \tag{33}$$

Let t be an integer such that $(t-1)/t \leqslant p_e \leqslant t/(t+1)$ and let α be a real number, $0 \leqslant \alpha \leqslant 1$, such that

$$p_e = \alpha \frac{t-1}{t} + (1-\alpha)\frac{t}{t+1}. \tag{34}$$

Then $\underline{U}(p_e)$ is attained if and only if

$$\Pr\left(\left\{x \;\middle|\; p(e \mid x) = \frac{t-1}{t}\right\}\right) = \alpha, \tag{35}$$

$$\Pr\left(\left\{x \;\middle|\; p(e \mid x) = \frac{t}{t+1}\right\}\right) = 1 - \alpha. \tag{36}$$

While $\overline{U}(p_e)$ can always be attained at any value of p_e, $\underline{U}(p_e)$ can be attained only for those values of p_e for which conditions (35) and (36) are met. If the features are continuous random variables then for any given values of p_e there exists a feature X with an appropriate density function such that conditions (35) and (36) are met. This is also true for binary features.

4. BOUNDS ON THE PROBABILITY OF ERROR BY MEANS OF f-ENTROPIES

The bounds $\overline{U}(p_e)$ and $\underline{U}(p_e)$ can also be used to derive upper and lower bounds on $P_X(e)$ for a given value of $U(X)$. Assume $U(X) = v$ and consider $\underline{p}_e$ and $\bar{p}_e$ as shown in Fig. 4. Then, clearly $\underline{p}_e \leqslant P_X(e) \leqslant \bar{p}_e$ (see Fig. 4). These bounds on $P_X(e)$ are not always explicit, but they can be computed by some known methods of numerical analysis.

Since there exist random variables which attain the upper and lower bound, $\bar{p}_e$, $\underline{p}_e$, these bounds are the tightest bounds which can be derived from f-entropies for an arbitrary set of features.

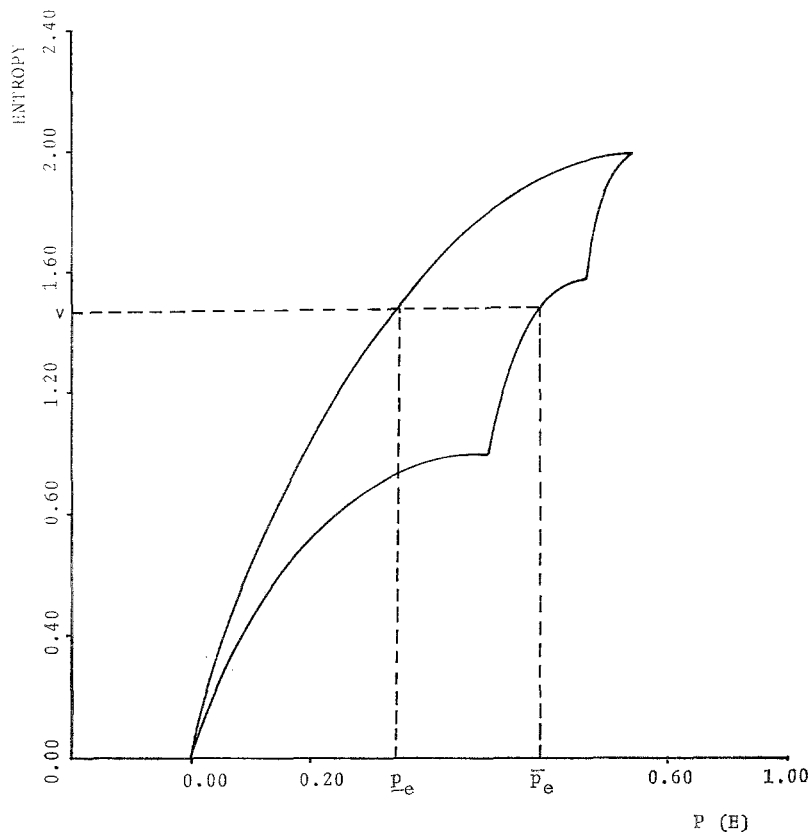

FIG. 4. Bounds on the probability of error by means of f-entropies.

EXAMPLE 1. $f_1(\pi) = \pi(1 - \pi)$. By (32)

$$\bar{U}_1(p_e) = (1 - p_e)\,p_e + (m - 1)\left(\frac{p_e}{m - 1}\right)\left(1 - \frac{p_e}{m - 1}\right),$$

which reduces to

$$\bar{U}_1(p_e) = -\frac{m}{m - 1}\,p_e^2 + 2p_e \tag{37}$$

By (33)

$$\underline{U}_1(p_e) = t\frac{1}{t}\left(1 - \frac{1}{t}\right) + t(t + 1)\left[(t + 1)\frac{1}{t + 1}\left(1 - \frac{1}{t + 1}\right) - t\frac{1}{t}\left(1 - \frac{1}{t}\right)\right]\left(p_e - \frac{t - 1}{t}\right) \tag{38}$$

which reduces to

$$\underline{U}_1(p_e) = p_e \,. \tag{39}$$

For a given random variable X, (37) and (39) imply

$$P_X(e) \leqslant U_1(X) \leqslant -\frac{m}{m-1} P_X^2(e) + 2P_X(e). \tag{40}$$

Solving the right-hand side inequality in (40) we obtain

$$P_X(e) \geqslant \frac{m-1}{m}\left[1 - \left(1 - \frac{m}{m-1} U_1(X)\right)^{1/2}\right]. \tag{41}$$

These bounds coincide with the results obtained by Cover and Hart (1967) regarding the relationship between the nearest neighbor risk and the Bayes risk. The same bounds have also been developed by Devijver (1973).

EXAMPLE 2. $f_2(\pi) = -\pi \log \pi$. This case has already been investigated by Kovalevski (1968) who found

$$\overline{U}(p_e) = -p_e \log p_e - (1 - p_e) \log(1 - p_e) + p_e \log(m - 1), \tag{42}$$

$$\underline{U}(p_e) = \log t + t(t+1) \log \frac{t+1}{t}\left(p_e - \frac{t-1}{t}\right), \tag{43}$$

where t is as defined in (28). The upper bound coincides with the well-known Fano bound (see Feinstein, 1958).

Kovalevski notes, that for the case $0 \leqslant p_e \leqslant \frac{1}{2}$, the parameter t attains the value 1 and then $\underline{U}(p_e)$ reduces to

$$\underline{U}(p_e) = 2p_e \,. \tag{44}$$

This implies the following upper bound

$$P_X(e) \leqslant \tfrac{1}{2} U_2(X) \tag{45}$$

which was derived later by Hellman and Raviv (1970) for any value of $P_X(e)$. However, for $P_X(e) > \frac{1}{2}$, the implicit bound (42) is a tighter bound than (45) (Chen, 1970).

EXAMPLE 3. $f_3(\pi) = A(\pi^\alpha - \pi)$, where $A = (2^{1-\alpha} - 1)^{-1}$. By (26) we obtain

$$\overline{U}(p_e) = A[(1 - p_e)^\alpha - (1 - p_e)] + A(m-1)\left[\left(\frac{p_e}{m-1}\right)^\alpha - \frac{p_e}{m-1}\right]$$

or

$$\overline{U}(p_e) = A\left[(1-p_e)^\alpha + (m-1)\left(\frac{p_e}{m-1}\right)^\alpha - 1\right]. \tag{46}$$

For $m = 2$ the bound in (46) coincides with the bound found by Toussaint (1974). In a forthcoming paper, Toussaint (1978) derives this bound for $m > 2$ as well.

By (29) we obtain

$$\underline{U}(p_e) = A[t^{1-\alpha} - 1] + A[t(t+1)^{2-\alpha} - (t+1)\,t^{2-\alpha}]\left[p_e - \frac{t-1}{t}\right]. \tag{47}$$

For $0 \leqslant p_e \leqslant \frac{1}{2}$ the parameter t attains the value 1 and then $\underline{U}(p_e)$ reduces to

$$\begin{aligned}\underline{U}(p_e) &= (2^{1-\alpha} - 1)^{-1}[2^{2-\alpha} - 2]\,p_e \\ &= 2p_e\end{aligned} \tag{48}$$

which implies the following upper bound

$$P_X(e) \leqslant \tfrac{1}{2}U_3(X) \qquad \text{for} \quad 0 \leqslant P_X(e) \leqslant \tfrac{1}{2}. \tag{49}$$

This explicit bound holds also for $P_X(e) > \frac{1}{2}$, however, for $P_X(e) > \frac{1}{2}$ the implicit bound (47) is tighter than (49).

5. Feature Selection by f-Entropies

THEOREM 8. *None of the members in T represents the relation $\gtrsim$ for an arbitrary set of features.*

Proof. Consider two features X and Y, for which $P_Y(e) = p_e$, $P_X(e) = p'_e$ and $p_e < p'_e$, i.e., $Y \succ X$. By Theorems 6 and 7, the upper and lower bounds $\underline{U}(p_e)$, $\overline{U}(p'_e)$ can be achieved for any given u, $u \in T$, and therefore, we may further assume that $U(Y) = \overline{U}(p_e)$ and $U(X) = \underline{U}(p'_e)$. Since $\overline{U}(p_e)$ and $\underline{U}(p_e)$ are continuous functions of p_e and for any given p_e, $\overline{U}(p_e) > \underline{U}(p_e)$ (except for $p_e = 1 - 1/m$ and perhaps $p_e = 0$, where equality holds) therefore, when p_e and p'_e are close enough, we obtain $\overline{U}(p_e) > \underline{U}(p'_e)$, which implies that U does not represent the relation $\gtrsim$. Figure 5 demonstrates the proof of the theorem. Q.E.D.

Many of the papers written on the subject of feature selection seem to be motivated by the feeling that there exists a magic functional which will induce the same ordering as does the probability of error. The significance of Theorem 8 is that such a functional, if it exists, does not belong to the f-entropy family.

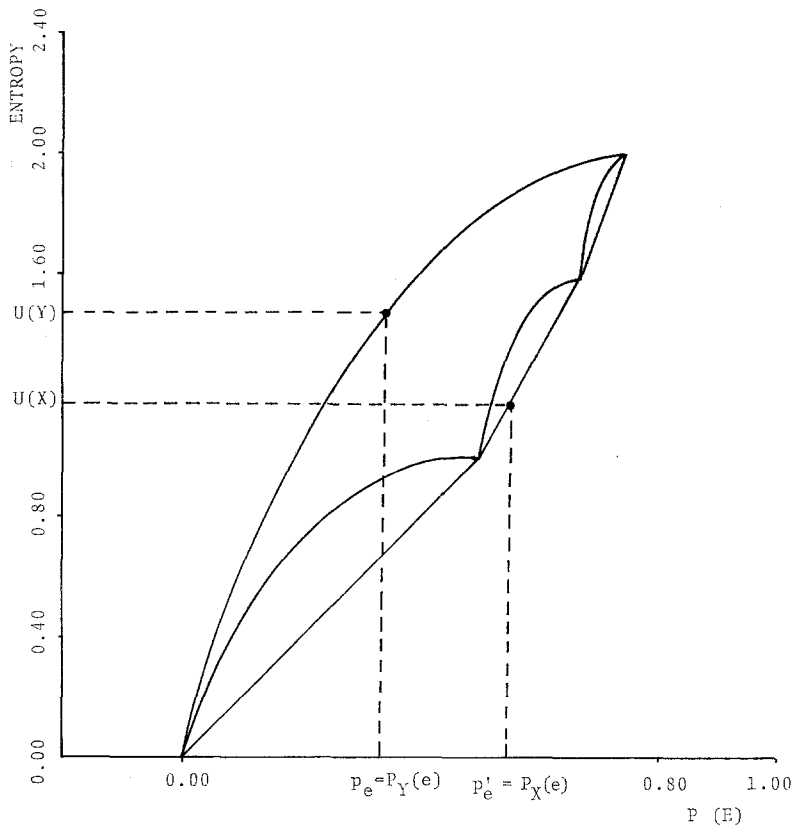

FIG. 5. $U(X) < U(Y)$ but $P_X(e) > P_Y(e)$.

In fact, the proof of Theorem 8 involves only the continuity of $\overline{U}(p_e)$ and $\underline{U}(p_e)$ as functions of p_e, and the strict inequality of the two functions for any given p_e, except, perhaps at the boundaries. It does not use any other properties which are unique to the f-entropy family, which implies that this theorem holds true for a wider family of feature selection rules. Proving Theorem 8 under the weakest possible conditions is beyond the scope of this paper.

Nevertheless, it is possible that for a specific set of features, one or more of the members in T does represent the relation $\gtrsim$. For a given set of features, it seems that the likelihood for a high correspondence between the ordering induced by the probability of error and the ordering induced by an f-entropy function u, $u \in T$, increases as the values attained by the difference function

$$\Delta(P_e) = \overline{U}(P_e) - \underline{U}(P_e) \tag{50}$$

decreases. In the extreme case, where $\Delta(p_e) = 0$ for $0 \leqslant p_e \leqslant 1 - 1/m$, $U(X)$ is

a monotone increasing function of $P_X(e)$, which obviously implies that $U(X)$ represents the relation $\gtrsim$. However, this extreme case cannot exist for u, $u \in T$, due to the strict concavity of $\bar{U}(p_e)$ and the convexity of $L(p_e)$.

The upper and lower bounds on the probability of error can be used for preliminary elimination of features which are dominated by other features.

THEOREM 9. *Let $\bar{p}_e$ be the upper bound for $P_Y(e)$ and let $\underline{p}_e$ be the lower bound for $P_X(e)$. If $\bar{p}_e \leqslant \underline{p}_e$ then $P_Y(e) \leqslant P_X(e)$, i.e., $Y \gtrsim X$.*

The proof of the theorem is an immediate conclusion from the following set of inequalities.

$$P_Y(e) \leqslant \bar{p}_e < \underline{p}_e \leqslant P_X(e). \tag{51}$$

This theorem holds for any pair of upper and lower bounds. However, the effectiveness of this elimination procedure increases as the tightness of the bounds increases.

6. SUMMARY

Characterization of the tightest lower and upper bounds on f-entropies by means of the probability of error have been presented. From those bounds the dual bounds on the probability of error by means of f-entropies have been derived. Using these bounds it has been proved that an f-entropy rule cannot induce on the set of all available features the same selection order as induced by the probability of error rule.

In a forthcoming paper, a comparison study between various members in the f-entropy family will be reported.

RECEIVED: May 21, 1976; REVISED: April 10, 1978

REFERENCES

1. BEN-BASSAT, M., AND RAVIV, J. (1976), Renyi's entropy, its properties and use in pattern recognition, presented at the Workshop on Pattern Recognition and Artificial Intelligence, Hyannis, June, 1976.
2. BEN-BASSAT, M., AND RAVIV, J. (1978), Renyi's entropy and the probability of error, *IEEE Trans. Inform. Theory*, in press.
3. BEN-BASSAT, M. (1978), ϵ-Equivalence of feature selection rules, *IEEE Trans. Inform. Theory*, in press.
4. BLACKWELL, D. (1951), Comparison of experiments, *in* "Proc. Second Berkeley Symp. on Probability and Statistics," Vol. I, pp. 93–102, Univ. of California Press, Berkeley.
5. BLACKWELL, D. (1953), Equivalent comparison of experiments, *Ann. Math. Stat.* **24**, 265–272.

6. CHEN, C. H. (1971), Theoretical comparison of a class of feature selection criteria in pattern recognition, *IEEE Trans. Comput.* **C20**, 1054–1056.
7. COVER, T. M., AND HART, P. E. (1967), Nearest neighbor pattern classification, *IEEE Trans. Inform. Theory* **IT13**, 21–27.
8. CSISZAR, I. (1967), Information type measures of difference of probability distributions and indirect observations, *Stud. Sci. Math. Hung.* **2**, 299–318.
9. DAROCZY, Z. (1970), Generalized information functions, *Inform. Contr.* **16**, 36–51.
10. DEGROOT, M. H. (1970), "Optimal Statistical Decisions," pp. 436–437, McGraw–Hill, New York.
11. DEVIJVER, P. A. (1973), On a new class of bounds on bayes risk in multihypothesis pattern recognition, *IEEE Trans. Comput.* **C23**, 70–80.
12. FEINSTEIN, A. (1958), "Foundations of Information Theory," McGraw–Hill, New York.
13. FERGUSON, T. S. (1967), "Mathematical Statistics," pp. 291–297, Academic Press, New York.
14. FU, K. S., MIN, P. J., AND LI, T. J. (1970), Feature selection in pattern recognition, *IEEE Trans. Systems Sci. Cybernetics* **SSC6**.
15. HELLMAN, M. E., AND RAVIV, J. (1970), Probability of error, equivocation and the Chernoff bound, *IEEE Trans. Inform. Theory* **IT16**, 368–372.
16. ITO, T. (1972), Approximate error bounds in pattern recognition, *in* "Machine Intelligence," Vol. VII, pp. 369–376, Edinburgh Univ. Press, Edinburgh, Scotland.
17. KAILATH, T. (1967), The divergence and Bhattacharyya distance in signal selection, *IEEE Trans. Commun. Technol.* **COM15**, 52–60.
18. KOVALEVSKI, V. A. (1968), The problem of character recognition from the point of view of mathematical statistics, *in* "Character Readers and Pattern Recognition" (V. A. Kovalevski, Ed.), pp. 3–30, Spartan Books, New York.
19. LISSACK, T., AND FU, K. S. (1976), Error estimation in pattern recognition via L^{α}-distance between posterior density functions, *IEEE Trans. Inform. Theory* **IT22**, 34–45.
20. MARILL, T., AND GREEN, D. M. (1963), On the effectiveness of receptors in recognition systems, *IEEE Trans. Inform. Theory* **IT9**, 11–17.
21. RENYI, A. (1960), On measures of entropy and information, *in* "Proc. Fourth Berkeley Symp. Math. Statist. and Probl., 1960," Vol. I, pp. 547–561, Univ. of California Press, Berkeley.
22. SHANNON, C. E. (1948), A mathematical theory of communication, *Bell Syst. Tech. J.* **27**, 379–423, 623–656.
23. TOUSSAINT, G. T. (1974), On information transmission, nonparametric classification and measuring dependence between random variables, *in* "Proceedings of the Symposium on Statistics and Related Topics," Carleton University, Canada.
24. TOUSSAINT, G. T. (1977), A generalization of Shannon's Equivocation and the Fano bound, *IEEE Trans. Systems, Man Cybernetics* **SMC7**, 300–302.
25. TOUSSAINT, G. T. (1978), Probability of error and equivocation of order α, *IEEE Trans. Inform. Theory*, in press.
26. VAJDA, I. (1968), Bounds on the minimal error probability and checking a finite or countable number of hypotheses, *Information Transmission Problems* **4**, 9–17.